Bilingual-Dictionary Adaptation to Domains

Hiroyuki Kaji
Central Research Laboratory, Hitachi, Ltd.
1-280 Higashi-Koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan
kaji@crl.hitachi.co.jp

Abstract

Two methods using comparable corpora to select translation equivalents appropriate to a domain were devised and evaluated. The first method ranks translation equivalents of a target word according to similarity of their contexts to that of the target word. The second method ranks translation equivalents according to the ratio of associated words that suggest them. An experiment using the EDR bilingual dictionary together with *Wall Street Journal* and *Nihon Keizai Shimbun* corpora proved that the method using the ratio of associated words outperforms the method based on contextual similarity. Namely, in a quantitative evaluation using pseudo words, the maximum F-measure of the former method was 86%, while that of the latter method was 82%.

1 Introduction

It is well known that appropriate translations for a word vary with domains, and bilingual-dictionary adaptation to domains is an effective way to improve the performance of, for example, machine translation and cross-language information retrieval. However, bilingual dictionaries have commonly been adapted to domains on the basis of lexicographers' intuition. It is thus desirable to develop an automated method for bilingual-dictionary adaptation.

Technologies for extracting pairs of translation equivalents from parallel corpora have been established (Gale and Church 1991; Dagan, et al. 1993; Fung 1995; Kitamura and Matsumoto 1996; Melamed 1997). They can, naturally, be used to adapt a bilingual dictionary to domains, that is, to select corpus-relevant translation equivalents from among those provided by an existing bilingual dictionary. However, their applicability is limited because of the limited availability of large parallel corpora. Methods of bilingual-dictionary adaptation using weakly comparable corpora, i.e., a pair of two language corpora of the same domain, are therefore required.

There are a number of previous works related to bilingual-dictionary adaptation using comparable corpora. Tanaka and Iwasaki's (1996) optimization method for a translation-probability matrix mainly aims at adapting a bilingual dictionary to domains. However, it is hampered by a huge amount of computation, and was only demonstrated in a small-scale experiment. Several researchers have developed a contextual-similarity-based method for extracting pairs of translation equivalents (Kaji and Aizono 1996; Fung and McKeown 1997; Fung and Yee 1998; Rapp 1999). It is computationally efficient compared to Tanaka and Iwasaki's method, but the precision of extracted translation equivalents is still not acceptable.

In the light of these works, the author proposes two methods for bilingual-dictionary adaptation. The first one is a variant of the contextual-similarity-based method for extracting pairs of translation equivalents; it focuses on selecting corpus-relevant translation equivalents from among those provided by a bilingual dictionary. This selecting may be easier than finding new pairs of translation equivalents. The second one is a newly devised method using the ratio of associated words that suggest each translation equivalent; it was inspired by a research on word-sense disambiguation using bilingual comparable corpora (Kaji and Morimoto 2002). The two methods were evaluated and compared by using the EDR (Japan Electronic Dictionary Research Institute) bilingual dictionary together with *Wall Street Journal* and *Nihon Keizai Shimbun* corpora.

2 Method based on contextual similarity

This method is based on the assumption that a word in a language and its translation equivalent in another language occur in similar contexts, albeit their contexts are represented by words in their respective languages. In the case of the present task (i.e., bilingual-dictionary adaptation), a bilingual dictionary provides a set of candidate translation equivalents for each target word[1]. The contextual similarity of each of the candidate translation equivalents to the target word is thus evaluated with the assistance of the bilingual dictionary, and a predetermined number of translation equivalents are selected in descending order of contextual similarity. Note that it is difficult to preset a threshold for contextual similarity since the distribution of contextual similarity values varies with target words.

[1] In this paper, "target word" is used to indicate the word for which translation equivalents are to be selected.

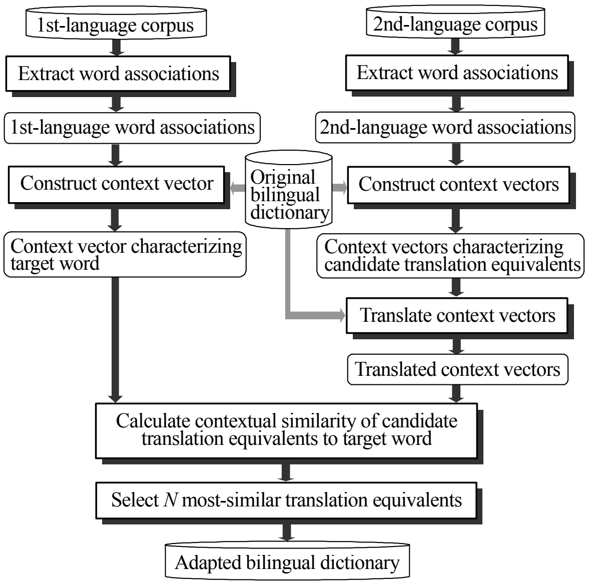
1st-language corpus
2nd-language corpus
Extract word associations
Extract word associations
1st-language word associations
2nd-language word associations
Construct context vector
Original bilingual dictionary
Construct context vectors
Context vector characterizing target word
Context vectors characterizing candidate translation equivalents
Translate context vectors
Translated context vectors
Calculate contextual similarity of candidate translation equivalents to target word
Select *N* most-similar translation equivalents
Adapted bilingual dictionary

Figure 1: Bilingual-dictionary adaptation based on contextual similarity

A flow diagram of the proposed method is shown in Figure 1. The essential issues regarding this method are described in the following.

Word associations are extracted by setting a threshold for mutual information between words in the same language. The mutual information of a pair of words is defined in terms of their co-occurrence frequency and respective occurrence frequencies (Church and Hanks 1990). A medium-sized window, i.e., a window including a few-dozen words, is used to count co-occurrence frequencies. Only word associations consisting of content words are extracted. This is because function words neither have domain-dependent translation equivalents nor represent contexts.

Both a target word and each of its candidate translation equivalents are characterized by context vectors. A context vector consists of associated words weighted with mutual information.

Similarity of a candidate translation equivalent to a target word is defined as the cosine coefficient between the context vector characterizing the target word and the translated context vector characterizing the candidate translation equivalent as follows. Under the assumption that target word x and candidate translation equivalent y are characterized by first-language context vector $\boldsymbol{a}(x) = (a_1(x), a_2(x), \ldots, a_\mathrm{m}(x))$ and second-language context vector $\boldsymbol{b}(y) = (b_1(y), b_2(y), \ldots, b_\mathrm{n}(y))$, respectively, $\boldsymbol{b}(y)$ is translated into a first-language vector denoted as $\boldsymbol{a'}(y) = (a'_1(y), a'_2(y), \ldots, a'_\mathrm{m}(y))$. That is,

$$a'_i(y) = \max_{j=1,2,\cdots,\mathrm{n}} \delta_{i,j} \cdot b_j(y) \quad (i = 1,2,\cdots,\mathrm{m}),$$

where $\delta_{i,j}$=1 if the j-th element of $\boldsymbol{b}(y)$ is a translation of the i-th element of $\boldsymbol{a}(x)$; otherwise, $\delta_{i,j}$=0. Elements of $\boldsymbol{b}(y)$ that cannot be translated into elements of $\boldsymbol{a'}(y)$ constitute a residual second-language vector, denoted as $\boldsymbol{b'}(y) = (b'_1(y), b'_2(y), \ldots, b'_\mathrm{n}(y))$. That is,

$$b'_j(y) = \begin{cases} b_j(y) & \cdots \quad \sum_{i=1}^{m} \delta_{i,j} = 0 \\ 0 & \cdots \quad \text{otherwise} \end{cases} \quad (j = 1,2,\cdots,\mathrm{n}).$$

The similarity of candidate translation equivalent y to target word x is then defined as

$$Sim(x,y) = cos(\boldsymbol{a}(x), \boldsymbol{a'}(y) + \boldsymbol{b'}(y)).$$

Note that $\boldsymbol{a'}(y)$+$\boldsymbol{b'}(y)$ is a concatenation of $\boldsymbol{a'}(y)$ and $\boldsymbol{b'}(y)$ since they have no elements in common.

3 Method using the ratio of associated words

3.1 Outline

This method is based on the assumption that each word associated with a target word suggests a specific sense of the target word, in other words, specific translation equivalents of the target word. It is also assumed that dominance of a translation equivalent in a domain correlates with how many associated words suggesting it occur in a corpus of the domain. It is thus necessary to identify which associated words suggest which translation equivalents. This can be done by using the sense-vs.-clue correlation algorithm that the author developed for unsupervised word-sense disambiguation (Kaji and Morimoto 2002). The algorithm works with a set of senses of a target word, each of which is defined as a set of synonymous translation equivalents, and it results in a correlation matrix of senses vs. clues (i.e., associated words). It is used here with a set of translation equivalents instead of a set of senses, resulting in a correlation matrix of translation equivalents vs. associated words.

The proposed method consists of the following steps (as shown in Figure 2).

First, word associations are extracted from a corpus of each language. The first step is the same as that of the contextual-similarity-based method described in Section 2.

Second, word associations are aligned translingually by consulting a bilingual dictionary, and pairwise correlation between translation equivalents of a target word and its associated words is calculated iteratively. A detailed description of this step is given in the following subsection.

Third, each associated word is assigned to the translation equivalent having the highest correlation with it. This procedure may be problematic, since an associated word often suggests two or more translation equivalents that represent the same sense. However, it is difficult to separate translation equivalents suggested by an associated word from others. Each associated word is therefore assigned to the translation equivalent it suggests most strongly.

Finally, a translation equivalent is selected when

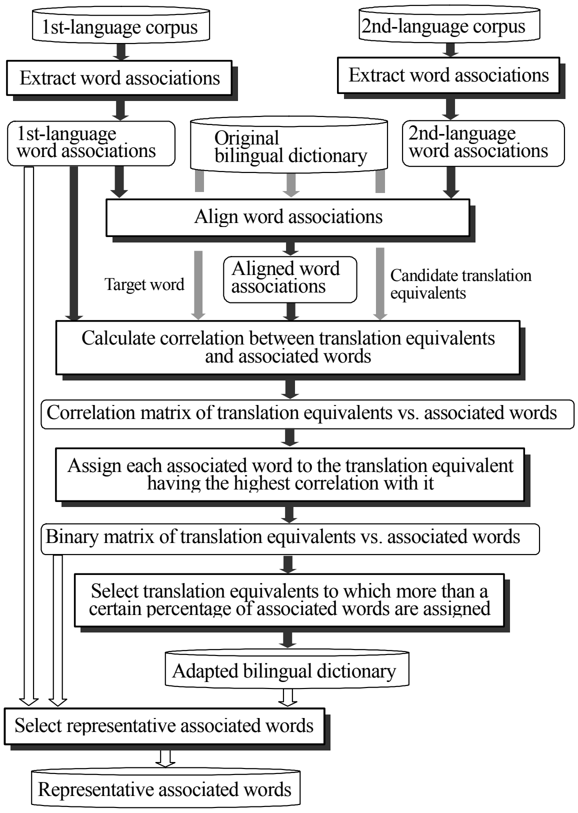
1st-language corpus
2nd-language corpus
Extract word associations
Extract word associations
1st-language word associations
Original bilingual dictionary
2nd-language word associations
Align word associations
Target word
Aligned word associations
Candidate translation equivalents
Calculate correlation between translation equivalents and associated words
Correlation matrix of translation equivalents vs. associated words
Assign each associated word to the translation equivalent having the highest correlation with it
Binary matrix of translation equivalents vs. associated words
Select translation equivalents to which more than a certain percentage of associated words are assigned
Adapted bilingual dictionary
Select representative associated words
Representative associated words

Figure 2: Bilingual-dictionary adaptation using the ratio of associated words

the ratio of associated words assigned to it exceeds a certain threshold. In addition, representative associated words are selected for each selected translation equivalent. A representativeness measure was devised under the assumption that representative associated words are near the centroid of a cluster consisting of associated words assigned to a translation equivalent. The representative associated words help lexicographers validate the selected translation equivalents.

3.2 Calculation of correlation between translation equivalents and associated words

The iterative algorithm described below has two main features. First, it overcomes the problem of failure in word-association alignment due to incompleteness of the bilingual dictionary and disparity in topical coverage between the corpora of the two languages. Second, it overcomes the problem of ambiguity in word-association alignment.

3.2.1 Alignment of word associations

For a first-language word association (x, $x'(j)$)—where a target word is given as x and its j-th associated word is given as $x'(j)$—a set consisting of second-language word associations alignable with it, denoted as $Y(x, x'(j))$, is constructed. That is,

$$Y(x, x'(j)) = \{(y, y') \mid (y, y') \in R_2, (x, y) \in D, (x'(j), y') \in D\},$$

where R_2 is the collection of word associations extracted from a corpus of the second language, and D is a bilingual dictionary to be adapted.

Each first-language word association (x, $x'(j)$) is characterized by a set consisting of accompanying associated words, denoted as $Z(x, x'(j))$. An accompanying associated word is a word that is associated with both words making up the word association in question. That is,

$$Z(x, x'(j)) = \{x'' \mid (x, x'') \in R_1, (x'(j), x'') \in R_1\},$$

where R_1 is the collection of word associations extracted from a corpus of the first language.

In addition, alignment of a first-language word association (x, $x'(j)$) with a second-language word association (y, y') ($\in Y(x, x'(j))$) is characterized by a set consisting of translingually alignable accompanying associated words, denoted as $W((x, x'(j)), (y, y'))$. A translingually alignable accompanying associated word is a word that is an accompanying associated word of the first-language word association making up the alignment in question and, at the same time, is alignable with an accompanying associated word of the second-language word association making up the alignment in question. That is,

$$W((x, x'(j)), (y, y')) = Z(x, x'(j)) \cap \{x'' \mid \exists y''(\in V(y, y'))\ (x'', y'') \in D\},$$

where $V(y, y') = \{y'' \mid (y, y'') \in R_2, (y', y'') \in R_2\}$.

3.2.2 Iterative calculation of correlation

The correlation between the i-th translation equivalent of target word x, denoted as $y(i)$, and the j-th associated word $x'(j)$ is defined as

$$C(y(i), x'(j)) = MI(x, x'(j)) \cdot \frac{PL(y(i), x'(j))}{\max_k PL(y(k), x'(j))},$$

where $MI(x, x'(j))$ is the mutual information between x and $x'(j)$, and $PL(y(i), x'(j))$ is the plausibility factor for $y(i)$ given by $x'(j)$. The mutual information between the target word and the associated word is the base of the correlation between each translation equivalent of the target word and the associated word; it is multiplied by the normalized plausibility factor. The plausibility factor is defined as the weighted sum of two component plausibility factors. That is,

$$PL(y(i), x'(j)) = PL_1(y(i), x'(j)) + \alpha \cdot PL_2(y(i), x'(j)),$$

where α is a parameter adjusting the relative weights of the component plausibility factors.

The first component plausibility factor, PL_1, is defined as the sum of correlations between the translation equivalent and the accompanying associated words. That is,

$$PL_1(y(i), x'(j)) = \sum_{x'' \in Z(x, x'(j))} C(y(i), x'').$$

This is based on the assumption that an associated word usually correlates closely with the translation equivalent that correlates closely with a majority of its accompanying associated words.

The second component plausibility factor, PL_2, is defined as the maximum plausibility of alignment involving the translation equivalent, where the plausibility of alignment of a first-language word association with a second-language word association is defined as the mutual information of the second-language word association multiplied by the sum of correlations between the translation equivalent and the translingually alignable accompanying associated words. That is,

$$PL_2\left(y(i),x'(j)\right)$$
$$= \max_{(y(i),y')\in Y(x,x'(j))}\left(MI(y(i),y')\cdot \sum_{x''\in W((x,x'(j)),(y(i),y'))} C\left(y(i),x''\right)\right).$$

This is based on the assumption that correct alignment of word associations is usually accompanied by many associated words that are alignable with each other as well as the assumption that alignment with a strong word association is preferable to alignment with a weak word association.

The above definition of the correlations between translation equivalents and associated words is recursive, so they can be calculated iteratively. Initial values are set as

$$C_0(y(i), x'(j)) = MI(x, x'(j)).$$

That is, the mutual information between the target word and an associated word is used as the initial value for the correlations between all translation equivalents of the target word and the associated word.

It was proved experimentally that the algorithm works well for a wide range of values of parameter α and that the correlation values converge rapidly. Parameter α and the number of iterations were set to five and six, respectively, in the experiments described in Section 4.

4 Experiments

4.1 Material and preparation

The experiment focused on nouns, whose appropriate translations often vary with domains. A wide-coverage bilingual noun dictionary was constructed by collecting pairs of nouns from the EDR English-to-Japanese and Japanese-to-English dictionaries. The resulting dictionary consists of 633,000 pairs of 269,000 English nouns and 276,000 Japanese nouns.

An English corpus consisting of *Wall Street Journal* articles (July 1994 to December 1995; 189MB) and a Japanese corpus consisting of *Nihon Keizai Shimbun* articles (December 1993 to November 1994; 275MB) were used as the comparable corpora. English nouns occurring 10 or more times in the English corpus were selected as the target words. The total number of selected target words was 12,848. For each target word, initial candidate translation equivalents were selected from the bilingual dictionary in descending order of frequency in the Japanese corpus; the maximum number of candidates was set at 20, and the minimum frequency was set at 10. The average number of candidate translation equivalents per target word was 3.3, and 1,251 target words had 10 or more candidate translation equivalents.

Extraction of word associations, which is the first step common to the method based on contextual similarity (abbreviated as the CS method hereinafter) and the method using the ratio of associated words (abbreviated as the RAW method hereinafter), was done as follows. Co-occurrence frequencies of noun pairs were counted by using a window of 13 words, excluding function words, and then noun pairs having mutual information larger than zero were extracted.

Table 1: Example translation equivalents selected by the method based on contextual similarity

Target word [Freq.]	#	Translation equivalent*) [Freq.]	Similarity
admini-stration [2027]	1	行政機関 (administration organ) [137]	0.127
	2	統治 (reign) [32]	0.119
	3	行政 (direction of domestic affairs) [2366]	0.116
	4	政権 (political power) [2370]	0.111
	5	施行 (operation) [453]	0.111
campaign [1656]	1	選挙運動 (election campaign) [71]	0.067
	2	競争 (competition) [2608]	0.050
	3	キャンペーン (aggressive activities) [561]	0.049
	4	運動 (movement) [947]	0.049
	5	軍事行動 (military activities) [89]	0.040
operation [3469]	1	経営 (management) [4810]	0.116
	2	事業 (enterprise) [8735]	0.091
	3	運営 (conduct) [1431]	0.076
	4	作戦 (tactics) [528]	0.074
	5	機能 (function) [2721]	0.074
power [2826]	1	エネルギー (energy) [913]	0.103
	2	力 (force) [6276]	0.101
	3	多数 (majority) [1036]	0.101
	4	電力 (electric power) [1208]	0.079
	5	能力 (ability) [1254]	0.074
shell [102]	1	殻 (husk) [135]	0.082
	2	球 (ball) [137]	0.070
	3	砲弾 (cannonball) [32]	0.070
	4	弾 (ball) [1370]	0.062
	6	ケース (case) [4851]	0.060
sign [4064]	1	声 (voice) [13536]	0.103
	2	目標 (target) [4676]	0.096
	3	景気 (business) [7163]	0.087
	4	兆候 (indication) [215]	0.087
	5	マーク (mark) [297]	0.084

*) English translations other than target words are given in parentheses.

4.2 Experimental results

Results of the CS and RAW methods for six target words are listed in Tables 1 and 2, respectively. Table 1 lists the top-five translation equivalents in descending order of contextual similarity. Table 2 lists translation equivalents with a ratio of associated words larger than 4% along with their top-four representative associated words. In these tables, the occurrence frequencies in the test corpora are appended to both the target words and the translation equivalents. These indicate the weak comparability between the *Wall Street Journal* and *Nihon Keizai Shimbun* corpora. Moreover, it is clear that neither the CS method nor the RAW method relies on the occurrence frequencies of words.

Tables 1 and 2 clearly show that the two methods produce significantly different lists of translation equivalents. It is difficult to judge the appropriateness of the results of the CS method without examining the comparable corpora. However, it seems that inappropriate translation equivalents were often ranked high by the CS method. In contrast, referring to the representative associated words enables the results of the RAW method to be judged as appropriate or inappropriate. More than 90% of the selected translation equivalents were judged as definitely appropriate.

Table 2 also includes the orders of translation equivalents determined by a conventional bilingual dictionary (remarks column). They are quite different from the orders determined by the RAW method. This shows the necessity and effectiveness of ranking translation equivalents according to relevancy to a domain.

Processing times were measured by separating both the CS and RAW methods into two parts. The processing time of the first part shared by the two methods, i.e., extracting word associations from corpora, is roughly proportional to the corpus size. For example, it took 2.80 hours on a Windows PC (CPU clock: 2.40 GHz; memory: 1 GB) to extract word associations from the 275 MB Japanese corpus. The second part, i.e., selecting translation equivalents for target words, is specific to each method, and the processing time of it is proportional to the number of target words. It took 11.5 minutes and 2.40 hours on another Windows PC (CPU clock: 2.40 GHz; memory: 512 MB) for the CS and RAW methods, respectively, to process the 12,848 target words. It was thus proved that both the CS and RAW methods are computationally feasible.

4.3 Quantitative evaluation using pseudo target words

4.3.1 Evaluation method

A method for bilingual-dictionary adaptation using comparable corpora should be evaluated by us-

Table 2: Example translation equivalents selected by the method using the ratio of associated words

Target word [Freq.]	#	Translation equivalent*) [Freq.]	Ratio	Representative associated words	Remarks **)
administration [2027]	1	内閣 (cabinet) [1067]	0.419	House, Clinton, White House, Republican	3a
	2	政権 (political power) [2370]	0.236	U.S. official, Haiti, Haitian, Clinton administration	3a
	3	施行 (operation) [453]	0.147	GATT, fast-track, trade pact, Trade	4a
	4	支配 (control) [84]	0.058	China, U.S., import, Japan	-
campaign [1656]	1	選挙運動 (election campaign) [71]	0.612	Republican, candidate, GOP, Democrat	2a
	2	キャンペーン (aggressive activities) [561]	0.371	ad, advertise, brand, advertising	2a
operation [3469]	1	経営 (management) [4810]	0.788	Stock Exchange, last year, profit, loss	2b
	2	事業 (enterprise) [8735]	0.144	quarter, net, income, plant	-
power [2826]	1	電力 (electric power) [1208]	0.434	electricity, power plant, utility, megawatt	8b
	2	勢力 (influence) [826]	0.425	military, leader, President, Haiti	3
	3	権限 (authority) [909]	0.062	reform, law, Ukraine, amendment	5a
shell [102]	1	砲弾 (cannonball) [32]	0.560	Serb, U.N., Sarajevo, NATO	4a
	2	貝 (shellfish) [100]	0.168	crab, fish, hermit crab, Mr. Soifer	1a
	3	球 (ball) [137]	0.112	rupture, bacterium, implant, brain	-
	4	外観 (external appearance) [267]	0.064	tape, camera, video, building	3a
sign [4064]	1	兆候 (indication) [215]	0.568	inflation, interest rate, rate, economy	4a
	2	看板 (signboard) [566]	0.099	tourist, billboard, airport, exit	3b
	3	目標 (target) [4676]	0.086	accord, agreement, pact, treaty	-
	4	兆し (indication) [2396]	0.086	last year, month, demand, order	-
	5	信号 (signal) [231]	0.062	driver, accident, highway, motorist	2a

*) English translations other than target words are given in parentheses.

**) This column shows the orders of translation equivalents determined by a conventional dictionary "Kenkyusha's New Collegiate English-Japanese Dictionary, 5th edition." For example, "3a" indicates that a translation equivalent belongs to the subgroup "a" in the third group of translations. A hyphen indicates that a translation equivalent is not contained in the dictionary.

ing recall and precision measures defined as

$$Recall = \frac{|S \cap T|}{|S|} \quad and \quad Precision = \frac{|S \cap T|}{|T|},$$

where S is a set consisting of pairs of translation equivalents contained in the test comparable corpora, and T is a set consisting of pairs of translation equivalents selected by the method. To calculate these measures, it is necessary to know all pairs of translation equivalents contained in the test corpora. This is almost impossible in the case that the test corpora are large.

To avoid this difficulty, an automated evaluation scheme using pseudo target words was devised. A pseudo word is formed by three real words, and it has three distinctive pseudo senses corresponding to the three constituent words. Translation equivalents of a constituent word are regarded as candidate translation equivalents of the pseudo word that represent the pseudo sense corresponding to the constituent word. For example, a pseudo word “action/address/application” has three pseudo senses corresponding to “action,” “address,” and “application.” It has candidate translation equivalents such as “訴訟<*SOSHOU*>” and “決議<*KETSUGI*>” originating from “action,” “演説<*ENZETSU*>” and “請願<*SEIGAN*>” originating from “address,” and “応用<*OUYOU*>” and “応募<*OUBO*>” originating from “application.” Furthermore, pseudo word associations are produced by combining a pseudo word with each of the associated words of the first two constituent words. It is thus assumed that first two pseudo senses occur in the corpora but the third one does not. For example, the pseudo word “action/address/application” has associated words including “court” and “vote,” which are associated with “action,” as well as “President” and “legislation,” which are associated with “address.”

Using the pseudo word associations, a bilingual-dictionary-adaptation method selects translation equivalents for the pseudo target word. On the one hand, when at least one of the translation equivalents originating from the first (second) constituent word is selected, it means that the first (second) pseudo sense is successfully selected. For example, when “訴訟<*SOSHOU*>” is selected as a translation equivalent for the pseudo target word “action/address/application,” it means that the pseudo sense corresponding to “action” is successfully selected. On the other hand, when at least one of translation equivalents originating from the third constituent word is selected, it means that the third pseudo sense is erroneously selected. For example, when “応用<*OUYOU*>” is selected as a translation equivalent for the pseudo target word “action/address/application,” it means that the pseudo sense corresponding to “application” is erroneously selected. The method is thus evaluated by recall and precision of selecting pseudo senses. That is,

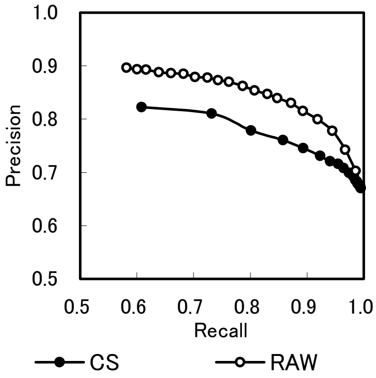

Figure 3: Recall and precision of selecting pseudo senses

$$Recall = \frac{|S' \cap T'|}{|S'|} \quad and \quad Precision = \frac{|S' \cap T'|}{|T'|},$$

where *S’* is a set consisting of pseudo senses corresponding to the first two constituent words, and *T’* is a set consisting of pseudo senses relevant to translation equivalents selected by the method.

4.3.2 Evaluation results

A total of 1,000 pseudo target words were formed by using randomly selected words that occur more than 100 times in the *Wall Street Journal* corpus. Using these pseudo target words, both the CS and RAW methods were evaluated. As for the CS method, the recall and precision of selecting pseudo senses were calculated in the case that N most-similar translation equivalents are selected (N=2, 3,…). As for the RAW method, the recall and precision of selecting pseudo senses were calculated in the case that the threshold for the ratio of associated words is set from 20% down to 1% in 1% intervals.

Recall vs. precision curves for the two methods are shown in Figure 3. These curves clearly show that the RAW method outperforms the CS method. The RAW method maximizes the F-measure, i.e., harmonic means of recall and precision, when the threshold for the ratio of associated words is set at 4%; the recall, precision, and F-measure are 92%, 80%, and 86%, respectively. In contrast, the CS method maximizes the F-measure when N is set at nine; the recall, precision, and F-measure are 96%, 72%, and 82%, respectively.

It should be mentioned that the above evaluation was done under strict conditions. That is, two out of three pseudo senses of each pseudo target word were assumed to occur in the corpus, while many real target words have only one sense in a specific domain. Target words with only one sense occurring in a corpus are generally easier to cope with than those with multiple senses occurring in a corpus. Accordingly, recall and precision for real target words would be higher than the above ones for the pseudo target words.

5 Discussion

The reasons for the superior performance of the RAW method to the CS method are discussed in the following.

- The RAW method overcomes both the sparseness of word-association data and the topical disparity between corpora of two languages. This is due to the smoothing effects of the iterative algorithm for calculating correlation between translation equivalents and associated words; namely, associated words are correlated with translation equivalents even if they fail to be aligned with their counterpart. In contrast, the CS method is much affected by the above-mentioned difficulties. All low values of contextual similarity (see Table 1) support this fact.
- The RAW method assumes that a target word has more than one sense, and, therefore, it is effective for polysemous target words. In contrast, contextual similarity is ineffective for a target word with two or more senses occurring in a corpus. The context vector characterizing such a word is a composite of context vectors characterizing respective senses; therefore, the context vector characterizing any candidate translation equivalent does not show very high similarity.
- The RAW method can select an appropriate number of translation equivalents for each target word by setting a threshold for the ratio of associated words. In contrast, the CS method is forced to select a fixed number of translation equivalents for all target words; it is difficult to predetermine a threshold for the contextual similarity, since the range of its values varies with target words (see Table 1).

Finally, from a practical point of view, advantages of the RAW method are discussed in the following.

- The RAW method selects translation equivalents contained in the comparable corpora of a domain together with evidence, i.e., representative associated words that suggest the selected translation equivalents. Accordingly, it allows lexicographers to check the appropriateness of selected translation equivalents efficiently.
- The ratio of associated words can be regarded as a rough approximation of a translation probability. Accordingly, a translation equivalent can be fixed for a word, when the particular translation equivalent has an exceedingly large ratio of associated words. A sophisticated procedure for word-sense disambiguation or translation-word selection needs to be applied only to words whose two or more translation equivalents have significant ratios of associated words.

6 Conclusion

The method using the ratio of associated words was proved to be effective, while the method based on contextual similarity was not. The former method has the following features that make it practical. First, is uses weakly comparable corpora, which are available in many domains. Second, it selects translation equivalents together with representative associated words that suggest them, enabling the translation equivalents to be validated. The method will be applied to several domains, and its effect on the performance of application systems will be evaluated.

7 Acknowledgments

This research was supported by the New Energy and Industrial Technology Development Organization of Japan (NEDO).

References

Church, Kenneth W. and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1): 22-29.

Dagan, Ido, Kenneth W. Church, and William A. Gale. 1993. Robust bilingual word alignment for machine aided translation. In *Proc. Workshop on Very Large Corpora*, pages 1-8.

Fung, Pascale. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proc. 33rd Annual Meeting of the ACL*, pages 236-243.

Fung, Pascale and Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. In *Proc. 5th Annual Workshop on Very Large Corpora*, pages 192-202.

Fung, Pascale and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proc. 36th Annual Meeting of the ACL / 17th COLING*, pages 414-420.

Gale, William A. and Kenneth W. Church. 1991. Identifying word correspondences in parallel texts. In *Proc. 4th DARPA Speech and Natural Language Workshop*, pages 152-157.

Kaji, Hiroyuki and Toshiko Aizono. 1996. Extracting word correspondences from bilingual corpora based on word co-occurrence information. In *Proc. 16th COLING*, pages 23-28.

Kaji, Hiroyuki and Yasutsugu Morimoto. 2002. Unsupervised word sense disambiguation using bilingual comparable corpora. In *Proc. 19th COLING*, pages 411-417.

Kitamura, Mihoko and Yuji Matsumoto. 1996. Automatic extraction of word sequence correspondences in parallel corpora, In *Proc. 4th Workshop on Very Large Corpora*, pages 79-87.

Melamed, I. Dan. 1997. A word-for-word model of translational equivalence. In *Proc. 35th Annual Meeting of the ACL / 8th Conference of the EACL*, pages 490-497.

Rapp, Reinhard. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proc. 37th Annual Meeting of the ACL*, pages 320-322.

Tanaka, Kumiko and Hideya Iwasaki. 1996. Extraction of lexical translations from non-aligned corpora, In *Proc. 16th COLING*, pages 580-585.